Next Token Generation & Word Embeddings

Nipun Batra

IIT Gandhinagar

July 31, 2025

Outline

Inspired by Excellence

This presentation is inspired by the excellent lecture from **Andrej Karpathy**

Inspired by Excellence

This presentation is inspired by the excellent lecture from **Andrej Karpathy**

Inspired by Excellence

This presentation is inspired by the excellent lecture from **Andrej Karpathy**

Find the original lecture:

Search for: "Neural Networks: Zero to Hero"

Inspired by Excellence

This presentation is inspired by the excellent lecture from **Andrej Karpathy**

Find the original lecture:

- Search for: "Neural Networks: Zero to Hero"
- Andrej Karpathy's educational content

Inspired by Excellence

This presentation is inspired by the excellent lecture from **Andrej Karpathy**

Find the original lecture:

- Search for: "Neural Networks: Zero to Hero"
- Andrej Karpathy's educational content

Inspired by Excellence

This presentation is inspired by the excellent lecture from **Andrej Karpathy**

Find the original lecture:

- Search for: "Neural Networks: Zero to Hero"
- Andrej Karpathy's educational content

Relevance to Modern Al

Connection to ChatGPT: The same fundamental principles we'll learn here power modern language models like GPT-4!

The Core Question

app

What is the next character?





Pose as Classification Task

We can frame this as a **multi-class classification problem**



Pose as Classification Task

We can frame this as a **multi-class classification problem**



Pose as Classification Task

We can frame this as a **multi-class classification problem**



Pose as Classification Task

We can frame this as a **multi-class classification problem**

Input:

app



Pose as Classification Task

We can frame this as a **multi-class classification problem**

Input:

Output: Probability Distribution

Character	Probability
а	0.01
b	0.01
С	0.03
 I	 0.85
•••	•••

Pop Quiz: Text Representation

Quick Quiz 1

Why can't we directly feed text into neural networks?

a) Text is too long for neural networks

Answer: b) Neural networks perform mathematical operations requiring numerical inputs!

Pop Quiz: Text Representation

Quick Quiz 1

Why can't we directly feed text into neural networks?

- a) Text is too long for neural networks
- b) Neural networks only work with numerical inputs

Answer: b) Neural networks perform mathematical operations requiring numerical inputs!

Pop Quiz: Text Representation

Quick Quiz 1

Why can't we directly feed text into neural networks?

- a) Text is too long for neural networks
- b) Neural networks only work with numerical inputs
- c) Text doesn't contain useful information

Answer: b) Neural networks perform mathematical operations requiring numerical inputs!

Our Training Dataset

Abid

- Abid
- Abhidha

- Abid
- Abhidha
- Adesh

- Abid
- Abhidha
- Adesh
- Ajay

- Abid
- Abhidha
- Adesh
- Ajay
- Akash

- Abid
- Abhidha
- Adesh
- Ajay
- Akash
- Ananya

- Abid
- Abhidha
- Adesh
- Ajay
- Akash
- Ananya
- •

- Abid
- Abhidha
- Adesh
- Ajay
- Akash
- Ananya
- •

Our Training Dataset

Abid

Priya

- Abhidha
- Adesh
- Ajay
- Akash
- Ananya
- •

- Abid
- Abhidha
- Adesh
- Ajay
- Akash
- Ananya
- •

- Priya
- Rahul

- Abid
- Abhidha
- Adesh
- Ajay
- Akash
- Ananya
- •

- Priya
- Rahul
- Sita

- Abid
- Abhidha
- Adesh
- Ajay
- Akash
- Ananya
- •

- Priya
- Rahul
- Sita
- Vikram

- Abid
- Abhidha
- Adesh
- Ajay
- Akash
- Ananya
- •

- Priya
- Rahul
- Sita
- Vikram
- :

- Abid
- Abhidha
- Adesh
- Ajay
- Akash
- Ananya
- •

- Priya
- Rahul
- Sita
- Vikram
- •
- Zara

- Abid
- Abhidha
- Adesh
- Ajay
- Akash
- Ananya
- •

- Priya
- Rahul
- Sita
- Vikram
- •
- Zara

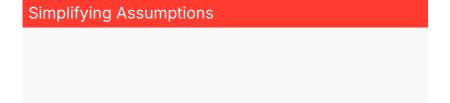
Our Training Dataset

- Abid
- Abhidha
- Adesh
- Ajay
- Akash
- Ananya
- •

- Priya
- Rahul
- Sita
- Vikram
- •
- Zara

Goal

Learn to generate new, realistic Indian names character by character



Simplifying Assumptions

1. Alphabet: Only 26 lowercase characters (a-z)

Simplifying Assumptions

1. Alphabet: Only 26 lowercase characters (a-z)

- 1. Alphabet: Only 26 lowercase characters (a-z)
- 2. End marker: Hyphen (-) indicates end of name

- 1. Alphabet: Only 26 lowercase characters (a-z)
- 2. End marker: Hyphen (-) indicates end of name

- 1. Alphabet: Only 26 lowercase characters (a-z)
- 2. End marker: Hyphen (-) indicates end of name
- 3. **Length:** Names are between 4-10 characters

- 1. Alphabet: Only 26 lowercase characters (a-z)
- 2. End marker: Hyphen (-) indicates end of name
- 3. **Length:** Names are between 4-10 characters

Simplifying Assumptions

- 1. Alphabet: Only 26 lowercase characters (a-z)
- 2. End marker: Hyphen (-) indicates end of name
- 3. Length: Names are between 4-10 characters

Example Encoding

"Abid" becomes "abid-"

From one name, create multiple training examples

From one name, create multiple training examples Example: "Abid" → "abid-"

From one name, create multiple training examples Example: "Abid" → "abid-"

Using Context/History of 3 Characters

X (Input)	Y (Target)
_	а
-a	b
-ab	i
abi	d
bid	-

From one name, create multiple training examples Example: "Abid" → "abid-"

Using Context/History of 3 Characters

X (Input)	Y (Target)
_	а
-a	b
-ab	i
abi	d
bid	-

From one name, create multiple training examples

Example: "Abid" → "abid-"

Using Context/History of 3 Characters

Y (Target)
а
b
i
d
-

Result: 5 training examples from one name "abid"!